DOCUMENT RESUME

ED 399 292 TM 025 605

AUTHOR Neill, Monty
TITLE How the "Principles and Indicators for Student Assessment Systems" Should Affect Practice.
INSTITUTION National Center for Fair and Open Testing (FairTest), Cambridge, MA.
PUB DATE Apr 96
NOTE 15p.; Version of a paper presented at the Annual Meeting of the American Educational Research Association (New York, NY, April 8-12, 1996).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Access to Education; *Educational Testing; Elementary Secondary Education; *Equal Education; Norm Referenced Tests; Objective Tests; *Performance Based Assessment; Psychological Testing; Psychometrics; *Standardized Tests; Standards; Test Bias; *Test Use
IDENTIFIERS *Principles and Indicators for Student Assessment; *Standards for Educational and Psychological Tests

ABSTRACT

The "Principles and Indicators for Student Assessment Systems" of the National Forum on Assessment, 1995, proposes a view of testing and assessment in elementary and secondary education that challenges the basic concepts and practices underlying the "Standards for Educational and Psychological Testing" of the American Educational Research Association and associated organizations. The "Standards," as they exist, are inadequate to the task of stopping the harmful social consequences of traditional standardized testing, but the "Principles" are constructed to place learning at the center of assessment. The basic model of educational testing addressed by the "Standards" relies on norm-referencing and on using multiple-choice or short-answer methods. Rather than enhancing access to education in the United States, the dominant forms of testing have limited access. In addition, they rely on outmoded psychological science. The seven "Principles" represent an agreement that traditional testing practices must change in the direction of becoming helpful for student learning. They replace the norm-referenced, multiple-choice short answer test with a complex of classroom-based assessments revolving around observation, documentation, and evaluation. They also assert that decisions about students must not be made on the basis of any single assessment. If the "Principles" were adopted in practice, the "Standards" would have to encourage more restrained use of tests and emphasize that assessment become compatible with what is known about human learning and development. (Contains 61 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 399 292

FairTest
National Center for Fair & Open Testing

How the *Principles and Indicators for Student Assessment Systems* Should Affect Practice

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Monty Neill, Ed.D.
Associate Director
National Center for Fair & Open Testing (FairTest)

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
MONTY NEILL
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Paper presented to the AERA Annual Meeting,
New York City, April 9, 1996
(Minor revisions, April 17, 1996)

The *Principles and Indicators for Student Assessment Systems* (National Forum on Assessment, 1995) proposes a view of testing and assessment in elementary and secondary education that challenges the basic concepts and practices underlying the *Standards for Educational and Psychological Testing* (American Educational Research Association, et al., 1985). I will argue here that traditional standardized testing in education has had predominantly harmful social consequences, and that the *Standards* are inadequate to the tasks of stopping the harmful consequences of testing or of ensuring that educational assessment performs what ought to be its primary task, enhancing learning for all students. The *Principles*, by contrast, are constructed to place support for learning at the center of assessment. I will draw out several implications for the practice of educational assessment and for the pending revision of the *Standards*.[1]

The first, fundamental question to ask of testing is what role, if any, should it play in society. That is, why test? By way of an answer, the Introduction to the *Standards* (AERA, et al., 1985, p.1) maintains, "Educational and psychological testing represents one of the most important contributions of behavioral science to our society...It has provided a tool for broader and more equitable access to education and employment." In other words, the document asserts that current forms of testing, including in education, have beneficial social consequences.

The Introduction does recognize that "testing has also been the target of extensive scrutiny, criticism, and debate," noting also, "The most frequent criticisms are that tests play too great a role in the lives of students and employees and that tests are biased and exclusionary" (p. 1). The *Standards*, however, never responds to these criticisms.

---

[1] I should note that while I am co-chair of the National Forum on Assessment, I am speaking here for FairTest, and my interpretation and use of the *Principles* does not necessarily represent the views of those who have signed the *Principles* or of other organizations that participate in the Forum.

BEST COPY AVAILABLE

TM025605

342 Broadway, Cambridge, Mass. 02139 (617) 864-4810 FAX (617) 497-2224

Instead, the document states that the *Standards* "is intended to provide a basis for evaluating the quality of testing practices as they affect the various parties involved" (p. 1). And though it is intended to "[e]mbody a strong ethical imperative," the *Standards* is "not a social action prescription" and does "not contain enforcement mechanisms" (p. *v*).

The *Standards*, we could then say, is simultaneously two things. First, it is a justification and defense of psychometrics, based on claims of science (testing is scientific) and beneficial consequences to social welfare (testing can make access more equitable and improve decision making). Second, it is a way of attempting to ensure the proper use of psychometric technology, thereby improving tests but also resolving or deflecting criticism.

Critique of Testing

Critics, including FairTest, remain unsatisfied. Their concerns are, if anything, stronger and broader than stated in the Introduction to the *Standards*. Indeed, critics have questioned the scientific underpinnings of testing since its earliest days; and they have charged that rather than expand access, testing has served to exclude, to deny or limit access, on the basis of class, race, gender and national origin.

The basic model of educational testing addressed by the *Standards* relies on norm-referencing and on using multiple-choice or short-answer methods (Gould, 1981; Resnick & Resnick, 1992; Taylor, 1994; Wiggins, 1993; Wolf, et al., 1991). Researchers have demonstrated that the scientific underpinnings of such testing, in particular the behavioral psychology on which it rests, are at best inadequate (Gardner, 1985; Resnick, 1987; Smith, 1986). The multiple-choice format of most educational testing has encouraged a view of learning that focuses on memorization, recognition and regurgitation of decontextualized bits of information (Frederiksen, 1984; Gardner, 1985; Resnick & Resnick, 1992; Smith, 1986). While this view of learning is strongly controverted by cognitive psychology (Gardner, 1985; Resnick, 1987; Smith, 1986), it lingers not only among test makers (Shepard, 1991a), but also among policymakers,[2] and no doubt among teachers and the general public. Unfortunately, a focus on memorization of isolated bits not only renders schooling dull, it is a method of instruction that simply fails to work for a great many students because it does not correspond with how people actually learn (Gardner, 1985; Resnick & Resnick, 1992; Smith, 1986). Multiple-choice is, however, the dominant method of testing (Garcia & Pearson, 1994).

Proponents of multiple-choice testing have garbed the method in the cloak of "objectivity." The simple response to this claim is that except for the scoring process, the tests are not objective: one or more subjective human beings decided everything, from what to test to how to test it, from writing items and choosing wanted answers and distractors to making decisions about the meaning of the results and how to use them. The very existence of

---

[2] For example, our work in Massachusetts has led us into dialogues with policymakers who have asserted that the foundations of learning, in particular reading, involve a process of acquiring decomposed pieces, labeled "basic skills," that can be measured one by one.

"objectivity" in the forms proposed by the philosophical positivism that underlies standardized tests has itself also been extensively challenged (e.g., Cherryholmes, 1988; Moss, 1996, 1992). Even if one accepts the positivist view of "objectivity" in philosophy, the fact remains that subjectivity is inescapable in assessment. More important, the educational consequences of this approach are not beneficial. As Johnston (1989) argues, the philosophy of science underlying testing presumes a model of education in which both teacher and student are objects, a view which disempowers both.

Norm-referencing in assessing educational achievement is a circular conception. It is justified primarily on evidence derived from the use of normal-curve tests and the social efforts to distribute opportunity and reward in a hierarchical manner (Bowles & Gintis, 1976; Taylor, 1994; Wolf, et al., 1991). Work based on norm-referencing may be technically sophisticated, but all that sophistication cannot overcome its circular presumptions. Norm-referencing reinforces the view that the ability to learn is distributed along the normal curve (Taylor, 1994; Wolf, et al., 1991). It thereby contributes to denying opportunities to students whose scores are low on the curve, often by narrowing the curriculum provided to those children (Allington, 1983; Bussis, 1982; Dorr-Bremme & Herman, 1986; Madaus, et al., 1992). Even most achievement tests are intended to compare students along a normal curve, not to determine how much and well students have learned what society has determined is important to learn (Taylor, 1994; Wolf et al., 1991; Neill & Medina, 1989; Wiggins, 1993).

As suggested above, researchers and critics have demonstrated that tests have served as gatekeepers, not gateways, for too many individuals, particularly from low-income, racial/ethnic minority, or recent immigrant groups, and women (Block & Dworkin, 1976; Kamin, 1977; Gould, 1981; Callahan, 1962; Bowles and Gintis, 1976; Karier, 1976; National Commission on Testing and Public Policy, 1990; Neill and Medina, 1989; Neill, 1993; Shepard and Smith, 1989). This gatekeeper effect involves entry into school (so-called "readiness tests"); placement in school in tracks or special programs, from "special education" to "gifted and talented"; grade promotion or retention ; graduation from high school; and entry into post-secondary education. Critics claim that testing narrows opportunities not only along various "demographic" lines, but also by unduly rewarding a narrow form of intellectual capability (Raven, 1992; Gardner, 1985). The use of testing to distribute rewards in ways that reinforce class and racial structures and to narrow and limit curriculum, means that testing, and by extension the *Standards*, has served to legitimate and perpetuate basic social inequities in the U.S.

Researchers have also well documented that testing has a strong impact on curriculum and instruction, so that testing determines not only what is and is not taught, but also how it is taught (Dorr-Bremme & Herman, 1986; Madaus, 1988; Madaus, et al., 1992; National Commission, 1990; Neill, 1993; Neill and Medina, 1989; Shepard, 1991b; Smith, 1991; Taylor, 1994; Wiggins, 1993; Wolf, et al., 1991). The effect of ceding control of curriculum and control of pedagogy to traditional standardized tests is demonstrably harmful. In substantial part, this problem stems from profound differences between the measurement

perspective and the instructional perspective.[3]

Students who do not come from "mainstream" families and who do not quickly grasp school culture and the dominant mode of teaching and learning, particularly memorization of decontextualized data and procedures, do not perform well on norm-referenced tests. The often-incorrect presumptions are then made that these students cannot learn well and that they need a stronger dose of what demonstrably has not worked (Oakes, 1985; Dentzer & Wheelock, 1990; Madaus, et al., 1992; National Commission, 1990; Shepard & Smith, 1989). Testing thus acts to determine the forms in which instruction and decision-making proceed, and then judge who does well by those forms. Unfortunately, the testing and instruction process is emotionally as well as intellectually stultifying (Raven, 1992). The damage is most severe to students from low-income and minority-group backgrounds, compounding the ways in which testing limits access.

I should add here that other forms of testing can have harmful consequences: criterion-referenced tests can actually incorporate norms and be used in similar fashion, and performance exams can be used to track, to deny opportunities, etc., and they may not assess cognitively complex learning or its application (Taylor, 1996; Messick, 1994). However, the dead hand of tradition enacted through the underlying paradigm of the multiple-choice, norm-referenced test that can be used as the basis for high stakes decisions should be understood as one of if not the primary obstacle to developing criterion or standards-referenced performance assessments (discussed below) that avoid the dangers discussed above.

In conclusion to this section, the dominant forms of educational testing and its primary uses in the U.S. are, regardless of the intentions of test makers and users, socially and educationally harmful, not helpful. Rather than enhance access, testing in the U.S. has limited access. Further, testing rests on what is at best outmoded psychological science. Thus, the two underpinnings of testing cited in the *Standards* -- that it is scientific and has beneficial consequences -- have been demonstrated to be false.

It is more accurate to refer to testing not as science but as a technology; and as Madaus (1994) has eloquently demonstrated, technologies, including testing, are not socially neutral. The evidence summarized above shows that the lack of neutrality is biased heavily against some groups in society, and that this lack of neutrality serves to sort and select students in ways that perpetuate the existing, often unfair, social order. Sorting and selecting are now, as they always have been, the primary purposes of testing in education, regardless of efforts to make testing more helpful and less biased. It is this underlying purpose, and the testing apparatus constructed to serve it, that is challenged by the *Principles and Indicators for Student Assessment Systems*.

---

[3] Indeed, these differences surfaced repeatedly in the writing of the *Principles*, and I believe contributed to some organizations not signing on to the document, which clearly favors instructional perspectives over measurement perspectives.

Principles and Indicators: Implications for Changing Practice

While the *Standards* is an ultimately unsuccessful effort to apply research and experience to the use of tests in a context in which testing is viewed as a positive social good, the *Principles* (National Forum, 1995) is an effort to apply research and experience to rethinking assessment in order to direct it toward the primary purpose of supporting student learning. It draws on the range of criticisms of traditional standardized testing (as noted above), knowledge thus far gained about the use of various forms of performance assessment (Berlak, et al., 1992; Darling-Hammond, et al., 1995; *Educational Leadership*, 1992, 1989; Estrin, 1993; Gardner, 1991; Linn, et al., 1991; Mathematical Sciences Education Board, 1993; McDonald, et al., 1993; Mitchell, 1992; National Council of Teachers of Mathematics, 1995; Neill, et al., 1995; Nettles & Nettles, 1995; Perrone, 1991; Valdez Pierce & O'Malley, 1992; Wiggins, 1993; Wolf, et al., 1991)[4]; research in a range of areas such as cognitive and developmental psychology (e.g., Gardner, 1985; Resnick, 1987; Smith, 1986); experience and knowledge from school reform efforts of the past decade, as shared by Forum members and others who participated in developing the *Principles*; and a shared vision of what schooling could and should be for all students. It is rooted in classroom and school experience of using assessment to support learning. It is deliberately what the *Standards* is not, a "social action prescription" (AERA, et al., 1985, p. *v*), though more in terms of defining a goal than describing how to attain the goal.

The *Principles*, developed collaboratively over a two-year period, has been signed by more than 80 national and regional education and civil rights organizations. It represents an agreement that 1) traditional testing practices must change, and 2) they must change in the direction of becoming helpful for student learning. The current primary impetus for testing -- sorting -- is instantly challenged by an approach that makes improving learning for all students primary.

Seven Principles

The document contains seven principles, as well as four "Educational Foundations for High Quality Assessment" which outline elements of schooling deemed essential by the Forum (see Appendix A for "Summary" of the *Principles*).[5] The Forum's principles are:

1. The primary purpose of assessment is to improve student learning.

2. Assessment for other purposes supports student learning.

---

[4] In addition to these general works, the *Principles* has a two-page bibliography "to provide readers with a general introduction to performance assessment" (p. 22-23). The knowledge in this field is expanding rapidly. See also FairTest (1995).

[5] A full copy of the *Principles* can be obtained from FairTest, 342 Broadway, Cambridge, MA 02139; $10.00.

3. Assessment systems are fair to all students.

4. Professional collaboration and development supports assessment.

5. The broad community participates in assessment development.

6. Communication about assessment is regular and clear.

7. Assessment systems are regularly reviewed and improved.

Assessment to support learning

Taken together, the first two principles clearly state the centrality of classroom assessments and the supportive role large-scale assessments must play. This presents a perspective which turns the current world of assessment on its head. For much of the past century, the model of assessment has been the on-demand, norm-referenced, multiple-choice test, the model which undergirds the *Standards*. With the *Principles*, the model becomes a set of rich, complex classroom practices, focusing on observation, documentation, and evaluation of actual student work done over time (see footnote 3).

In this new paradigm, assessment is interwoven with curriculum and instruction, not just something that happens after the fact. It requires teachers to use a variety of forms and methods. It encourages multiple ways for students to demonstrate their learning, and it provides students with opportunities to actively apply knowledge through projects, exhibitions, performances, and portfolios, as well as exams. The model also promotes student choice and self-evaluation, individual and group work, and continuous feedback to students. Multiple-choice and short-answer methods, and assessments constructed to sort or rank-order students (particularly norm-referenced tests), if used at all, constitute only a limited part of the total assessment system. Thus, that which is fundamental to the sorts of testing focused on by the *Standards* is pushed to the margins, and that which has been marginal is made central.

To work well, such assessment presumes both high-quality curriculum and equity for all students. Believing that all students can learn to high levels, the Forum recommends that "Schools establish clear statements of desired learning for all students and help all students achieve them." Such standards "describe broad, important intellectual competencies -- knowledge, skills, understandings, and habits of mind -- that students should acquire and be able to demonstrate." Thus, the *Principles* focus on assessments geared toward standards of learning rather than toward normative comparisons.

In order to assist classroom learning, assessments must be able to indicate individual development as a thinker and doer, or to be what Johnston termed "self-referenced" (Johnston, 1992; see also, Carini, 1994). Additionally, such assessments must be "theory referenced" (Johnston, 1992); that is, rooted in theories of learning, of cognition, and of the

domains, that are appropriately rich and well-developed (Johnston, 1992; Neill, et al., 1995; Resnick & Resnick, 1992). Put another way, the behavioral psychology undergirding traditional tests needs replacing by improved psychological theory, which the *Principles* calls for in its "Foundations" section when it says, "Schools work to understand how learning takes place and what facilitates learning" (National Forum, 1995, p. 4). In effect, the *Principles* seeks to rely on cognitive and sociocultural understandings of learning and human development (Nelson-Barber & Trumbull Estrin, n.d.), urging that such knowledge be used in developing assessments compatible with learning.

That classroom-based assessment involves subjectivity is not disputed, but subjectivity is seen as an asset, not a problem. As humans necessarily are involved in evaluation in education, the key issue is to improve the capability of the human assessors, not to try to eliminate them by misleading beliefs in objectivity (see Principle 4).

In sum, the *Principles* replaces the norm-referenced, multiple-choice/short answer test with a complex of classroom-based assessments revolving around observation, documentation and evaluation. In this process, the instructional uses of assessment take precedence over other uses, and thus the conceptions used to shape assessment necessarily change from those of measurement to those of teaching. Technical issues important to assessment and measurement do not disappear, but they must respond to changed priorities. As the *Principles* puts it, "Technical standards for assessment are revised or developed to ensure they are adequate for the assessment purposes and methods, and they are used to help ensure high quality practices."

Improvement and Accountability

The *Principles* proposes basic changes in using assessment data for making decisions about students, planning school improvement, and ensuring accountability to the public. Instead of relying primarily on one-time standardized exams, even performance exams, the Forum recommends relying primarily on evidence of learning collected in the classroom over time for all these purposes.

The *Principles* states that decisions about students, such as high school graduation or grade promotion, should not be made on the basis of any single assessment. This is in sharp contrast to the *Standards*, which effectively presume the regular use of one-time tests for making decisions, though the *Standards* does maintain, with regard to educational testing, that "a decision or characterization that will have a major impact on a test taker would not automatically be made on the basis of a single test score" (Standard 8.12, p. 54). This Standard should be expanded and strengthened in the forthcoming revision. It is a good case of a Standard often ignored, as well as a good case for which enforcement, at least at the level of public censure for the many states and districts that make high stakes decisions solely on the basis of tests, would be a great help.

Assessment for school improvement should rely primarily on information gathered in the school about student work over time. In their book, *Authentic Assessment in Action*, Darling-

Hammond, Ancess and Falk (1995) show how five schools of various kinds are using performance assessments to make decisions about students and to improve education, from changing curriculum to rethinking the structure of the school day. Essentially, the assessments provide rich data for use in thinking about improvement. In addition, the processes of doing classroom assessment and using the resulting information help create an environment of thoughtful reflection on how to improve curriculum and instruction. Again, the kinds of assessments used flow from an instructional perspective rather than from a measurement view.

For district and state accountability information, the *Principles* recommend "a combination of classroom assessment information (such as portfolio reviews) and external or large-scale assessments (such as examinations)" (Principle 2, p. 8). Sampling should be used to the extent feasible.

Relying on teacher evaluation for a major part of accountability data introduces some technical difficulties. However, the principle of using grades -- which are based on an evaluation of student work done over time -- to determine high school graduation is widely accepted socially, legally, and politically, even though it is also widely agreed that teacher grading usually lacks technical rigor. (Here it is worth reminding the reader that despite all the variability in grades, they are more accurate predictors of performance in the first year of college than are the technically rigorous SAT or ACT; see College Board, 1995). In effect, the *Principles* proposes that strengthened evaluation by teachers become an important basis for accountability. This, in turn, calls for an improved form of "grading," preferably without the numbers or letters or the competitive rankings (Kohn, 1994).

The involvement of parents and the community in the assessment process, discussed in Principles 5 and 6, also enhances accountability. This requires that assessment be open, not cloaked in the traditionally prized secrecy (see Principle 1). As Wiggins (1993) explains, secrecy operates to make education deeply dishonest, undermining what ought to be important goals of learning. This is not a call for parents to score their own child's work, but for involving the community in a variety of ways, from working on learning goals to participating in such things as reviewing exhibitions or performances (Darling-Hammond, et al., 1995).

One might ask, in regard to overall assessment practices, why not combine both approaches, classroom-based assessment and traditional tests, which are admittedly inexpensive? This approach has been argued for under the rubric of "multiple measures," or at times as a call to not "throw the baby out with the bathwater." I hope I have explained why the traditional tests are not simply inadequate but also harmful; that there is no baby in the bathwater. The continuing social weight of those tests also means that their continued use, even in combination with other assessments, will tell educators that they can keep right on focusing instruction on what traditional tests measure. Multiple measures are of course necessary, but the term does not mean one of those measures must be a traditional test.

Equity and Professional Development

Traditional tests have presumed that assessing all students in the same format creates an equitable situation. However, the process of test construction, the determination of content, and the use of only one method -- usually multiple-choice -- build in cultural and educational biases that unfairly favor some ways of understanding and demonstrating knowledge over others. Testing's power has, in turn, shaped curriculum and instruction in ways that favor certain groups. Norm-referenced testing has encouraged often-harmful educational practices, such as tracking (see discussion above). Thus, the uniformity and apparent equity of the tests contribute to real world educational inequity.

The *Standards* functionally defines bias only in terms of predictive validity. It explicitly avoids the issue of "fairness" (p. 13). This ignores the multiple and complex ways in which bias can affect all aspects of the assessment process. For example, in developing an exam, bias must be avoided in developing the framework for the construct, in defining the domain, in selecting items or tasks meant to assess student knowledge in that domain, and in specifying outcome criteria against which a test is validated. Yet these issues are largely ignored in the *Standards*.

Bias also has existed in classroom assessment. For example, teachers may be inequitable in scoring and evaluation, unfairly rewarding some ways of demonstrating knowledge and some people over others. Accountability must therefore also serve equity.

When accountability is based on classroom information, there will be a set of back-up documents that can be examined. For example, if Latino children in a particular school or district generally do not score as well as White/Anglo children, an investigative team could look at the portfolios, work samples, etc., on which the scores are based. School practice can thus be held up to scrutiny, as has been done with portfolios in Pittsburgh (LeMahieu, Gitomer & Eresh, 1995; see also, Neill, et al., 1995).

Improvement in teacher assessment practices also can help ensure equity. If teachers really know how to look at each individual child, to know her strengths and ways of learning, his cultural background and interests, then they can work better and more fairly with each student. Professional development, therefore, should include a focus on using assessment with a diverse student body.

Additionally, professional development should help teachers better understand different roads to high quality outcomes. For example, through discussions which center on reviewing student work, teachers can improve their knowledge of students, confront their biases, and learn how to work better with their students. In this process, they strengthen the school as a community of learners. Thus assessment becomes part of school improvement and a means for increasing equity, two important elements of accountability (Darling-Hammond, Ancess & Falk, 1995; Neill, et al., 1995). This approach is certainly counter to that which insists on only one way to demonstrate knowledge, usually in a format that can only assess well-constructed problems with one "correct" answer. The one-right-answer approach is, in Norm

Fredericksen's (1984) words, "the real test bias," because most important problems are ill-structured and have more than one reasonably correct response.

Implications for the *Standards*

Making the enhancement of student learning central to assessment thinking and practice; prioritizing classroom assessment and thereby changing the paradigmatic model of assessment away from norm-referenced, multiple-choice tests; ceasing to make decisions based on one-time events; focusing on helping all students meet high but varied standards rather than ranking for sorting -- these are the changes in assessment practice called for by the *Principles*. But, as argued at the start, these principles are radically different from those that propelled the development of testing in the U. S. and which undergird and structure the *Standards*.

If the *Principles* were adopted in practice, much of the *Standards'* focus would change. The role of technical standards, in general, as well as the concerns over the need for enforcement, also would change. In closing, then, let me briefly consider these issues.

The *Standards* are largely a set of guidelines for preparing and using the kinds of tests that have virtually no legitimate role in education. While guidelines for assessment practices should include technical standards, the current *Standards* is not an adequate document in light of the *Principles*. If the educational research community takes seriously the need to make assessment serve learning, the AERA should not support a revision of the *Standards* that is anything less than a profound transformation.

The concern for enforcement, a concern shared by FairTest, arises primarily from the kinds of tests used in the kinds of ways that ought to be eradicated. If the *Principles* is followed, then concerns such as whether a test-taker's rights were respected when her or his score is questioned (Haney, 1996) are moot. However, so long as some scarce goods are to be distributed on the basis of prior achievement, there is a need to ensure that the determination be fair and accurate. Technical standards do have a role to play in this process, and enforcing such standards will remain an issue. Thus, the AERA should take steps to insist that some form of enforcement be developed. If the other sponsoring organizations do not wish to develop such a process, the AERA should proceed on its own to do so.

In conclusion, to be compatible with the *Principles*, the *Standards* will have to encourage a more restrained use of tests and powerfully emphasize that assessment become compatible with what is known about human learning and development as well as a far richer appreciation of academic content than has traditionally been the case. Assessment must be constructed on a stronger scientific basis. Issues of fair assessment in a complex and diverse society cannot be reduced to predictive calculations. Norm-referencing and multiple-choice testing must no longer be used to narrow classroom assessment, never mind curriculum and instruction. Rather, assessment must be used to improve learning and opportunities for all students.

Monty Neill is Associate Director of the National Center for Fair & Open Testing (FairTest) and co-chair of the National Forum on Assessment. He can be reached at 342 Broadway, Cambridge, MA 02139; e-mail to mneillft@aol.com.

Bibliography

Allington, R.L. (1983). The reading instruction provided readers of differing reading abilities. *Elementary School Journal, 83*(5), 548-559.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Berlak, H., Newmann, F. M., Adams, E., Archbald, D. A., Burgess, T., Raven, J., & Romberg, T. A. (Eds.). (1992). *Toward a new science of educational testing & assessment*. Albany, NY: State University of New York Press.

Bussis, A.M. (December, 1982). "Burn it at the casket": Reading instruction and children's learning of the first R. *Phi Delta Kappan*, pp. 237-241.

Block, N.J., & Dworkin, G. (1976). *The IQ controversy: Critical Readings*. New York: Pantheon.

Bowles, S., & Gintis, H. (1976). *Schooling in capitalist America*. New York: Basic Books.

Callahan, R. (1962). *Education and the cult of efficiency*. Chicago: University of Chicago Press.

Carini, P. F. (1994). Dear Sister Bess: An Essay on Standards, Judgement and Writing. *Assessing Writing, 1*(1), 29-65.

Cherryholmes, C. H. (1988). Construct validity and discourses of research. *American Journal of Education, 96*(3), 421-457.

College Board. (1995). *Counselor's handbook for the SAT program*. Author.

Darling-Hammond, L., Ancess, J., & Falk, B. (1995). *Authentic assessment in action: Studies of schools and students at work*. New York: Teachers College Press.

Dentzer, E., & Wheelock, A. (1990). *Locked in/locked out: Tracking and placement practices in Boston public schools*. Boston: Massachusetts Advocacy Center.

Dorr-Bremme, D.W., & Herman, J.L. (1986). *Assessing student achievement: A profile of classroom practices*. CSE monograph series in evaluation, 11. Los Angeles: Center for the Study of Evaluation.

Educational Leadership. (1989). Special Issue: Redirecting Assessment. 46(7).

Educational Leadership. (1992). Special Issue: Using Performance Assessment. 49(8).

Estrin, E. T. (1993). Alternative assessment: Issues in language, culture, and equity. *Knowledge Brief,* #11. San Francisco: Far West Laboratory.

FairTest. (1995). *Performance assessment: Annotated bibliography and resources -- revised.* Cambridge, MA: National Center for Fair & Open Testing (FairTest).

Frederiksen, N. (March 1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist, 39*, pp. 193-202.

Garcia, G.E., & Pearson, P.D. (1994). Assessment and diversity. In L. Darling-Hammond (Ed.), *Review of research in education, 20* (pp. 337-391). Washington, DC: American Educational Research Association.

Gardner, H. (1991). Assessment in context: The alternative to standardized testing. In B. Gifford & M.C. O'Connor, (Eds.), *Cognitive approaches to assessment.* Boston: Kluwer Academic.

Gardner, H. (1985). *The mind's new science*. New York: Basic Books.

Gould, S.J. (1981). *The mismeasure of man.* New York: Norton.

Haney, W. (April, 1996). Standards, schmandards: The need for bringing test standards to bear on assessment practice. New York: Paper presented at the American Educational Research Association annual meeting.

Johnston, P. H. (1992). *Constructive evaluation of literate activity*. New York: Longman.

Johnston, P.H. (1989). Constructive evaluation and the improvement of teaching and learning. *Teachers College Record, 90*(4).

Kamin, L. (1977). The politics of IQ. In P.L. Houts (Ed.), *The myth of measurability*. New York: Hart.

Karier, C.J. (1976). Testing for order and control in the corporate liberal state. In Block & Dworkin (Eds.).

Kohn, A. (October 1994).Grading: The issue is not how but why. *Educational Leadership, 52*(2), 38-41.

LeMahieu, P., D. Gitomer, and J. Eresh. (Fall 1995). "Portfolios in Large-Scale Assessment: Difficult But Not Impossible." *Educational Measurement, 14*(3).

Linn, R.L, Baker, E.L., & Dunbar, S.B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.

Madaus, G. (1994). A technological and historical consideration of equity issues associated with proposals to change our nation's testing policy. *Harvard Educational Review, 64*(1), 76-95.

Madaus, G. F. (1988). The influence of testing on the curriculum. *87th Yearbook of the national society for the study of education, Part I: Critical issues in the curriculum,* 83-121.

Madaus, G. F., West, M. M., Harmon, M. C., Lomax, R. G., & Viator, K. A. (1992). *The influence of testing on teaching math and science in grades 4-12 (SPA8954759).* Chestnut Hill, MA: Boston College, Center for the Study of Testing, Evaluation, and Educational Policy.

Mathematical Sciences Education Board. (1993). *Measuring what counts: A conceptual guide for mathematics assessment.* Washington, DC: National Academy Press.

McDonald, J. P., Smith, S., Turner, D., Finney, M., & Barton, E. (1993). *Graduation by exhibition: Assessing genuine achievement.* Alexandria, VA: Association for Supervision and Curriculum Development.

Messick, S. (1994) The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23,* 13-23.

Mitchell, R. (1992). *Testing for learning: How new approaches to evaluation can improve American schools.* New York: Free Press.

Moss, P. A. (1996). Enlarging the dialogue in educational measurement: Voices from interpretive research traditions. *Educational Researcher, 25*(1), 20-28, 43.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62*(3), 229-258.

National Commission on Testing and Public Policy. (1990). *From gatekeeper to gateway: Transforming testing in America.* Chestnut Hill, MA: Author.

National Council of Teachers of Mathematics (NCTM). (1995). *Assessment standards for school mathematics.* Reston, VA: Author.

National Forum on Assessment. (1995.) *Principles and Indicators for Student Assessment Systems.* Cambridge, MA: FairTest.

Neill, D. M. (1993). Standardized testing: Harmful to civil rights. In United States Commission on Civil Rights, *The validity of testing in education and employment.* Washington, DC: Author.

Neill, M., Bursh, P., Schaeffer, R., Thall, C., Yohe, M., & Zappardino, P. (1995). *Implementing performance assessment: A guide to classroom, school and system reform.* Cambridge, MA: National Center for Fair & Open Testing (FairTest).

Neill, M., & Medina, N. J. (1989). Standardized testing: Harmful to educational health. *Phi Delta Kappan, 70,* 688-697.

Nelson-Barber, S., & Trumbull Estrin, E. (n.d. - 1995-6). *Culturally responsive mathematics and science education for native students.* San Francisco: Far West Laboratory.

Nettles, M. T., & Nettles, A. L. (Eds.). (1995). *Equity and excellence in educational testing and assessment*. Norwell, MA: Kluwer Academic Publishers.

Oakes, J. (1985). *Keeping track: How schools structure inequality*. New Haven, CT: Yale University Press.

Perrone, V. (Ed.). (1991). *Expanding student assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.

Raven, J. (1992). A model of competence, motivation, and behavior, and a paradigm for assessment. In H. Berlak, et al. (Eds.), *Toward a new science of educational testing and assessment* (pp. 85-116). Albany, NY: State University of New York Press.

Resnick, L. B. (1987). *Education and learning to think*. Washington, DC: National Academy Press.

Resnick, L. B. & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Future assessments: Changing views of aptitude, achievement, and instruction*. Boston: Kluwer.

Shepard, L. A. (1991a) Will national tests improve student learning? *Phi Delta Kappan, 73,* 232-238.

Shepard, L. A. (1991b).Psychometricians' beliefs about learning. *Educational Researcher, 20* (6), 2-16.

Shepard, L.A., & Smith, M.L. (1989). *Flunking grades: Research and policies on retention*. Philadelphia: Falmer Press.

Smith, F. (1986). *Insult to intelligence: The bureaucratic invasion of our classrooms*. New York: Arbor House.

Smith, M.L. (1991). Put to the test: The effects of external testing on teachers. *Educational Researcher, 20*(5), 8-11.

Taylor, C. (Summer 1994). Assessment for Measurement or Standards: The Peril and Promise of Large-Scale Assessment reform. *American Educational Research Journal, 31*(2), 231-262.

Valdez Pierce, L., & O'Malley, J.M. (1992). *Performance and portfolio assessment for language minority students*. Washington, DC: National Clearinghouse for Bilingual Education.

Wiggins, G. P. (1993). *Assessing Student Performance: Exploring the Purpose and Limits of Testing*. San Francisco: Jossey-Bass Publishers.

Wolf, D., Bixby, J., Glenn, J., III, & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. In G. Grant (Ed.), *Review of research in education, 17,* (pp. 31-74). Washington, DC: American Educational Research Association.